Time series model based on global structure of complete genome

Zu-Guo Yu[1,2] and Vo Anh[1*]

[1]Centre for Statistical Science and Industrial Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, Q 4001, Australia.

[2]Department of Mathematics, Xiangtan University, Hunan 411105, P. R. China.[†]

Abstract

A time series model based on the global structure of the complete genome is proposed. Three kinds of length sequences of the complete genome are considered. The correlation dimensions and Hurst exponents of the length sequences are calculated. Using these two exponents, some interesting results related to the problem of classification and evolution relationship of bacteria are obtained.

PACS numbers: 87.10+e, 47.53+n

Key words: Correlation dimension, Hurst exponent, Coding/noncoding segments, complete genome,

1 Introduction

The nucleotide sequences stored in GenBank have exceeded hundreds of millions of bases and they increase by ten times every five years. A great deal of information concerning origin of life, evolution of species, development of individuals, and expression and regulation of genes, exist in these sequences[1]. In the past decade or so there has been an enormous interest in unravelling the mysteries of DNA. It has become very important to improve on new theoretical methods to do DNA sequence analysis. Statistical analysis of DNA sequences[1−9] using modern statistical measures is proven to be particularly fruitful. There is another approach to research DNA, namely nonlinear scales method, such as fractal dimension[10, 11, 12, 13], complexity[14, 16]. The correlation properties of coding and noncoding DNA sequences was first studied by Stanley and coworkers[5] in their "fractal landscape or DNA walk" model. The DNA walk defined in [5] is that the

*E-mail, Zu-Guo Yu: yuzg@hotmail.com or z.yu@qut.edu.au, Vo Anh: v.anh@qut.edu.au

[†]This is the permanent corresponding address of Zu-Guo Yu.

walker steps "up" if a pyrimidine (C or T) occurs at position i along the DNA chain, while the walker steps "down" if a purine (A or G) occurs at position i. Stanley and coworkers[5] discovered there exists long-range correlation in noncoding DNA sequences while the coding sequences correspond to regular random walk. But if one considers more details by distinguishing C from T in pyrimidine, and A from G in purine (such as two or three dimensional DNA walk model[1] and maps given in [14]), then the presence of base correlation has been found even in coding region. However, DNA sequences are more complicated than those these types of analysis can describe. Therefore, it is crucial to develop new tools for analysis with a view toward uncovering mechanisms used to code other types of information.

Since the first complete genome of the free-living bacterium *Mycoplasma genitalium* was sequenced in 1995[17], an ever-growing number of complete genomes has been deposited in public databases. The availability of complete genomes opens the possibility to ask some global questions on these sequences. The avoided and under-represented strings in some bacterial complete genomes have been discussed in [13, 18, 19]. A time series model of CDS in complete genome has also been proposed in [15].

One can ignore the composition of the four kind of bases in coding and noncoding segments and only consider the roughly structure of the complete genome or long DNA sequences. Provata and Almirantis [20] proposed a fractal Cantor pattern of DNA. They map coding segments to filled regions and noncoding segments to empty regions of random Cantor set and then calculate the fractal dimension of the random fractal set. They found that the coding/noncoding partition in DNA sequences of lower organisms is homogeneous-like, while in the higher eucariotes the partition is fractal. This result is interesting and reasonable, but it seems too rough to distinguish bacteria because the fractal dimensions of bacteria they gave out are all the same. The classification and evolution relationship of bacteria is one of the most important problem in DNA research. In this paper, we propose a time series model based on the global structure of the complete genome and we find that one can get more information from this model than that of the fractal Cantor pattern. We have found some new results to the problem of classification and evolution relationship of bacteria.

A DNA sequence is a sequence over the alphabet $\{A, C, G, T\}$ representing the four bases from which DNA is assembled, namely adenine, cytosine, guanine, and thymine. But from views of the level of structure, the complete genome of organism is made up of coding and noncoding segments. Here the length of a coding/noncoding segment means the number of its bases. First we simply count out the lengths of coding/noncoding segments in the complete genome. Then we can get three kinds of integer sequences by the following ways.

i) First we order all lengths of coding and noncoding segments according to the order of coding and noncoding segments in the complete genome, then replace the lengths of noncoding segments by their negative numbers. So that we can distinguish lengths of coding and noncoding segments. This integer sequence is named *whole length sequence*.

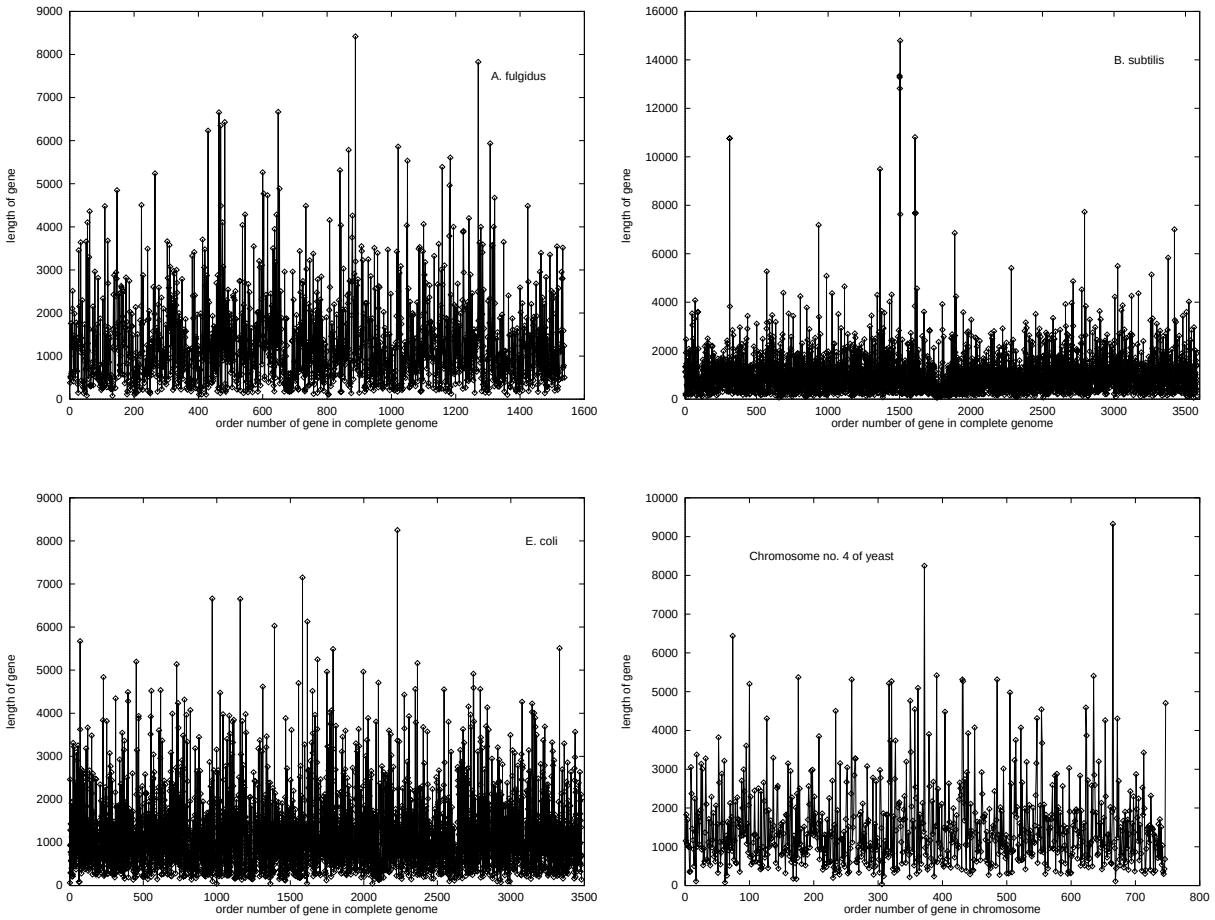

Figure 1: Length and distribution of coding segments in the complete genome or Chromosome of some organisms.

ii) We order all lengths of coding segments according to the order of coding segments in the complete genome. We name this integer sequence *coding length sequence*. For some examples, we plot the distribution of coding length sequences of three bacteria genome and the 4th chromosome of Saccharomyces cerevisiae (yeast) in Figure 1.

iii) We order all lengths of noncoding segments according to the order of noncoding segments in the complete genome. This integer sequence is named *noncoding length sequence*.

We can now view these three kinds of integer sequences as time series. We want to calculate their correlation dimensions and Hurst exponents.

2 Correlation dimension and Hurst exponent

The notion of correlation dimension, introduced by Grassberger and Procaccia[21, 22], suits well experimental situations, when only a single time series is available. It is now being used widely in many branches of physical science. Consider a sequence of data from

a computer or laboratory experiment:

$$x_1,\ x_2,\ x_3, \cdots,\ x_N, \tag{1}$$

where N is a large enough number. These numbers are usually sampled at an equal time interval $\Delta\tau$. We embed the time series into $\mathbf{R}^m$, choose a time delay $\tau = p\Delta\tau$, then obtain

$$\mathbf{y}_i = (x_i, x_{i+p}, x_{i+2p}, \cdots, x_{i+(m-1)p}), \quad i = 1, 2, \cdots, N_m, \tag{2}$$

where

$$N_m = N - (m-1)p. \tag{3}$$

In this way we get N_m vectors in the embedding space $\mathbf{R}^m$.

For any $\mathbf{y}_i, \mathbf{y}_j$, we define the distance as

$$r_{ij} = d(\mathbf{y}_i, \mathbf{y}_j) = \sum_{l=0}^{m-1} |x_{i+lp} - x_{j+lp}|. \tag{4}$$

If the distance is less than a given number r, we say that these two vectors are correlated. The correlation integral is defined as

$$C_m(r) = \frac{1}{N_m^2} \sum_{i,j=1}^{N_m} H(r - r_{ij}), \tag{5}$$

where H is the Heaviside function

$$H(x) = \begin{cases} 1, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases} \tag{6}$$

For a proper choice of m and not too big a value of r, it has been shown by Grassberger and Procaccia[22] that the correlation integral $C_m(r)$ behaves like

$$C_m(r)\ \propto\ r^{D_2(m)}. \tag{7}$$

Thus one can define the correlation dimension as

$$D_2 = \lim_{m \longrightarrow \infty} D_2(m) = \lim_{m \longrightarrow \infty} \lim_{r \longrightarrow 0} \frac{\ln C_m(r)}{\ln r}. \tag{8}$$

For more details on D_2, the reader can refer to [23].

To deal with practical problems, one usually choose $p = 1$. If we choose a sequence $\{r_i :\ 1 \leq i \leq n\}$ such that $r_1 < r_2 < r_3 < \cdots < r_n$, then a scaling region can be found in the $\ln r - \ln C_m(r)$ plane, see [23], p.346. Then the slop of the scaling region is $D_2(m)$. When $D_2(m)$ does not change with m increasing, we can take this $D_2(m_0)$ as the estimate value of D_2. We calculate the correlation dimensions of three kinds of length sequences of the complete genome using the method introduced above. From the $\ln r - \ln C_m(r)$ figures of these sequences of different values of embedding dimension m, we find that it is

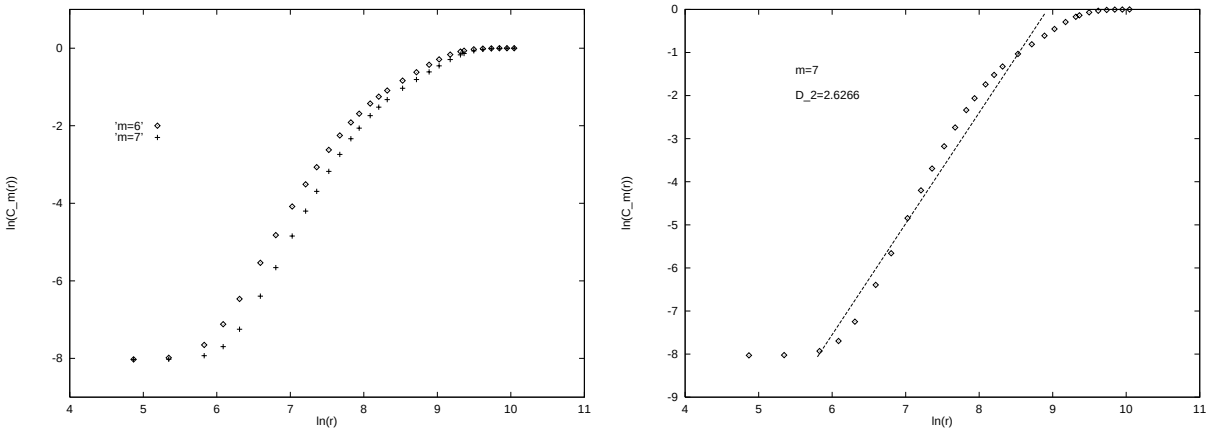

Figure 2: Left) $\ln r$-$\ln C_m(r)$ figure of the length sequence of coding and noncoding segments of A. fulgidus when m=6,7. Right) Estimate of the correlation dimension (the continuous line).

suitable to choose $m = 7$. For example, we give the $\ln r - \ln C_m(r)$ figure of whole length sequence of A. fulgidus when $m = 6, 7$ (Figure 2). We take the region from the third point to the 20th point (from left to right) as the scaling region.

Hurst[24] invented the now famous statistical method — *the rescaled range analysis* (R/S analysis) to study the long-range dependence in time series. Later on, B. B. Mandelbrot[25] and J. Feder [26] brought R/S analysis into fractal analysis. For any time series $x = \{x_k\}_{k=1}^N$ and any $2 \leq n \leq N$, one can define

$$< x >_n = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{9}$$

$$X(i,n) = \sum_{u=1}^{i} [x_u - < x >_n] \tag{10}$$

$$R(n) = \max_{1 \leq i \leq n} X(i,n) - \min_{1 \leq i \leq n} X(i,n) \tag{11}$$

$$S(n) = [\frac{1}{n} \sum_{i=1}^{n} (x_i - < x >_n)^2]^{1/2}. \tag{12}$$

Hurst found that

$$R(n)/S(n) \ \sim \ (\frac{n}{2})^H. \tag{13}$$

H is called the *Hurst exponent*.

As n changes from 2 to N, we obtain $N-1$ points in the $\ln(n)$ v.s. $\ln(R(n)/S(n))$ plane. Then we can calculate the Hurst exponent H of the length sequence of organisms using the least-squares linear fit. As an example, we plot the graph of R/S analysis of the whole length sequence of A. fulgidus in Figure 3.

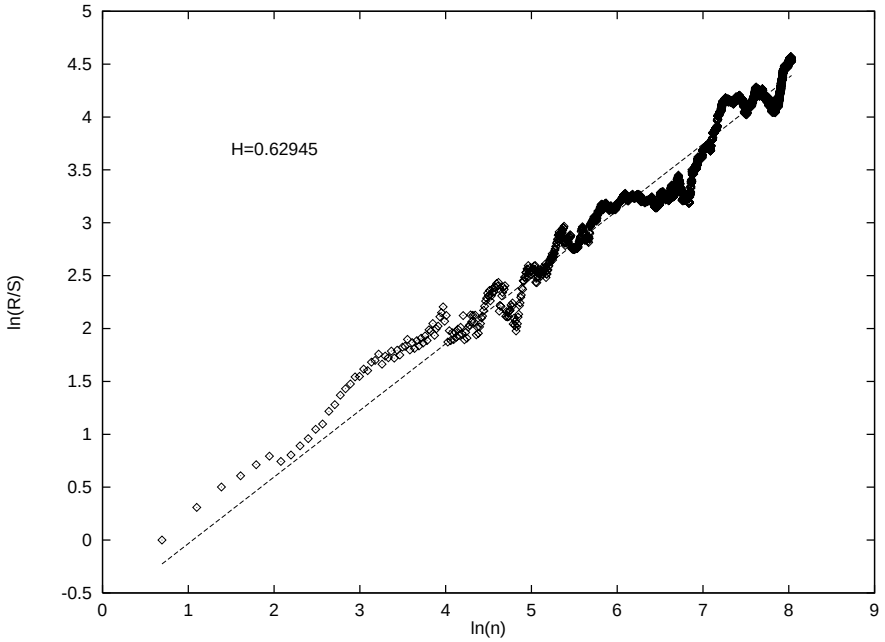

Figure 3: Calculation of Hurst exponent.

The Hurst exponent is usually used as a measure of complexity. The trajectory of the record is a curve with fractal dimension $D = 2 - H$ ([25],p.149). Hence a smaller H means a more complex system. When applied to fractional Brownian motion, the system is said to be *persistent* if $H > 1/2$, which means that if for a given time period t, the motion is along one direction, then in a succeeding time, it is more likely that the motion will follow the same direction. For $H < 1/2$, the opposite holds, that is, the system is *antipersistent*. But when $H = 1/2$, the system is a Brownian motion, and is random.

3 Data and results.

More than 21 bacterial complete genomes are now available in public databases . There are five Archaebacteria: Archaeoglobus fulgidus, Pyrococcus abyssi, Methanococcus jannaschii, Aeropyrum pernix and Methanobacterium thermoautotrophicum; four Gram-positive Eubacteria: Mycobacterium tuberculosis, Mycoplasma pneumoniae, Mycoplasma genitalium, and Bacillus subtilis. The others are Gram-negative Eubacteria. These consist of two Hyperthermophilic bacteria: Aquifex aeolicus and Thermotoga maritima; six proteobacteria: Rhizobium sp. NGR234, Escherichia coli, Haemophilus influenzae, Helicobacter pylori J99, Helicobacter pylori 26695 and Rockettsia prowazekii; two chlamydia Chlamydia trachomatis and Chlamydia pneumoniae, and two Spirochete: Borrelia burgdorferi and Treponema pallidum.

We calculate the correlation dimensions and Hurst exponents of three kinds of length sequences of the above 21 bacteria. The estimated results are given in Table 1 (we denote

Table 1: $D_{2,whole}$, $D_{2,cod}$ and $D_{2,noncod}$ of 21 bacteria.

$D_{2,whole}$	$D_{2,cod}$	$D_{2,noncod}$	Species of Bacterium	Category
2.1126	1.3581	1.1612	Mycoplasma genitalium	Gram-positive Eubacteria
2.3552	1.7102	1.5077	Mycoplasma pneumoniae	Gram-positive Eubacteria
2.5239	1.8891	0.8944	Aquifex aeolicus	Hyperthermophilic bacteria
2.5125	1.9094	0.5849	Thermotoga maritima	Hyperthermophilic bacteria
2.2705	2.0119	2.2014	Rhizobium sp. NGR234	Proteobacteria
2.4060	2.0378	0.4695	Borrelia burgdorferi	Spirochete
2.4561	2.0729	0.6145	Treponema pallidum	Spirochete
2.5345	2.1674	1.3001	Chlamydia trachomatis	Chlamydia
2.6015	2.3055	1.3187	Chlamydia pneumoniae	Chlamydia
2.6096	2.4137	0.8475	Pyrococcus abyssi	Archaebacteria
2.5617	2.4589	2.1515	Rickettsia prowazekii	Proteobacteria
2.6266	2.4867	0.7011	Archaeoglobus fulgidus	Archaebacteria
2.6916	2.5195	1.2134	Aeropyrum pernix	Archaebacteria
2.6497	2.5248	0.9239	Helicobacter pylori 26695	Proteobacteria
2.6353	2.5364	0.9555	Helicobacter pylori J99	Proteobacteria
2.7181	2.8417	1.1126	Haemophilus influenzae	Proteobacteria
2.6558	2.8861	1.1427	Methanococcus jannaschii	Archaebacteria
2.5687	2.9097	0.6862	M. thermoautotrophicum	Archaebacteria
2.8473	2.9250	1.1031	Mycobacterium tuberculosis	Gram-positive Eubacteria
2.8984	3.0976	1.3660	Escherichia coli	Proteobacteria
2.7039	3.2435	1.1035	Bacillus subtilis	Gram-positive Eubacteria

by $D_{2,whole}$, $D_{2,cod}$ and $D_{2,noncod}$ the correlation dimensions of whole, coding and noncoding length sequences, from top to bottom, in the increasing order of the value of $D_{2,cod}$) and Table 2 (we denote by H_{whole}, H_{cod} and H_{noncod} the Hurst exponents of whole, coding and noncoding length sequences, from top to bottom, in the increasing order of the value of H_{cod}).

4 Discussion and conclusions

Although the existence of the archaebacterial urkingdom has been accepted by many biologists, the classification of bacteria is still a matter of controversy[27]. The evolutionary relationship of the three primary kingdoms (i.e. archeabacteria, eubacteria and eukaryote) is another crucial problem that remains unresolved[27].

From Table 1, we can roughly divide bacteria into two classes, one class with $D_{2,cod}$ less

Table 2: H_{whole},H_{cod} and H_{noncod} of 21 bacteria.

H_{whole}	H_{cod}	H_{noncod}	Species of Bacterium	Category
0.3904	0.3311	0.6446	Rhizobium sp. NGR234	Proteobacteria
0.4280	0.4108	0.5640	Pyrococcus abyssi	Archaebacteria
0.4063	0.4381	0.5925	Rickettsia prowazekii	Proteobacteria
0.4736	0.4660	0.5504	Helicobacter pylori 26695	Proteobacteria
0.4828	0.5147	0.4648	Mycoplasma genitalium	Gram-positive Eubacteria
0.5064	0.5343	0.5381	Chlamydia pneumoniae	Chlamydia
0.5979	0.5365	0.5873	Helicobacter pylori J99	Proteobacteria
0.4731	0.5445	0.6005	Chlamydia trachomatis	Chlamydia
0.5297	0.5698	0.5626	Mycobacterium tuberculosis	Gram-positive Eubacteria
0.5410	0.5882	0.4948	Thermotoga maritima	Hyperthermophilic bacteria
0.5288	0.5941	0.6843	Mycoplasma pneumoniae	Gram-positive Eubacteria
0.5362	0.5985	0.4655	Escherichia coli	Proteobacteria
0.5528	0.6017	0.3153	M. thermoautotrophicum	Archaebacteria
0.6295	0.6098	0.6311	Archaeoglobus fulgidus	Archaebacteria
0.6013	0.6145	0.4605	Aquifex aeolicus	Hyperthermophilic bacteria
0.5202	0.6153	0.5136	Haemophilus influenzae	Proteobacteria
0.5727	0.6371	0.4986	Aeropyrum pernix	Archaebacteria
0.6830	0.6622	0.6764	Borrelia burgdorferi	Spirochete
0.7213	0.6894	0.5612	Treponema pallidum	Spirochete
0.7271	0.7183	0.6399	Bacillus subtilis	Gram-positive Eubacteria
0.7732	0.7793	0.3607	Methanococcus jannaschii	Archaebacteria

than 2.40, and the other with $D_{2,cod}$ greater than 2.40. We observe that the classification of bacteria using $D_{2,cod}$ almost coincides with the traditional classification of bacteria. All Archaebacteria belong to the same class. All Proteobacteria belong to the same class except Rhizobium sp. NGR234, in particular, the closest Proteobacteria Helicobacter pylori 26695 and Helicobacter pylori J99 group with each other. Two Spirochete group with each other. Two Chlamydia gather with each other. Gram-positive bacteria is divided into two sub-categories: Mycoplasma genitalium and Mycoplasma pneumoniae belong to one class and gather with each other, Mycobacterium tuberculosis and Bacillus subtilis belong to another class and almost gather with each other.

If one classifies bacteria using $D_{2,whole}$, with the $D_{2,whole}$ of one subclass less than 2.55, that of the other larger than 2.55, almost the same results hold as those using $D_{2,cod}$. But when one classifies bacteria using $D_{2,noncod}$, the results are quite different. This is quite reasonable because the coding segments occupy the main part of space of the DNA chain of bacteria.

A surprising feature shown in Table 1 is that the Hyperthermophilic bacteria (including Aquifex aeolicus and Thermotoga maritima) are linked closely with the Archaebacteria if we only consider the length sequences of noncoding segments. But when we consider the length sequences of coding segments, they are linked closely with eubacteria. We notice that Aquifex, like most Archaebacteria, is hyperthermophilic. Hence it seems that their hyperthermophilicity property is possibly controlled by the noncoding part of the genome, contrary to the traditional view resulting from classification based on the coding part of the genome. It has previously been shown that Aquifex has close relationship with Archaebacteria from the gene comparison of an enzyme needed for the synthesis of the amino acid trytophan[28]. Such strong correlation on the level of complete genome between Aquifex and Archaebacteria is not easily accounted for by lateral transfer and other accidental events[28]. Our result is based on different levels of the genome from that used by the authors of [28]

From Table 1, one can also see the $D_{2,cod}$ values are almost larger than the $D_{2,noncod}$ values. Hence the coding length sequences are more complex than the noncoding length sequences.

From Table 2, we can also roughly divide bacteria into two classes, one class with H_{cod} less than 0.60, and the other with H_{cod} greater than 0.60. One can see all Archeabacteria belong to the same class except Pyrococcus abyssi. All Gram-positive Eubacteria belong to the same class except Bacillus subtilis. All Proteobacteria belong to the same class except Haemophilus influenzae. Two Spirochete group with each other. Two Chlamydia almost group with each other.

We also find the H_{noncod} values of all Archeabacteria except Pyrococcus abyssi, two Hyperthermophilic bacteria, and Mycoplasma genitalium and E. coli are less than 1/2, while those of other bacteria are greater than 1/2. Hence Hyperthermophilic bacteria have some common information with Archaebacteria in noncoding segments.

We calculate $D_{2,whole}$, $D_{2,cod}$, $D_{2,noncod}$ H_{whole}, H_{cod} and H_{noncod} of the 4th chromosome

of Saccharomyces cerevisiae (yeast). They are 2.5603, 2.1064, 2.5013, 0.5517, 0.6255 and 0.5482 respectively. From Tables 1 and 2, if we consider $D_{2,whole}$, H_{whole} and H_{cod}, we can see that Archaebacteria and Chlamydia are linked more closely with yeast which belongs to eukaryote than other categories of bacteria. There are several reports (such as [29]) that, in some RNA and protein species, archeabacteria are much more similar in sequences to eukaryotes than to eubacteria. Our present result supports this point of view.

In [14], we find that the Hurst exponent is a good tool to distinguish different functional regions. But now considering more global structure of the genome, we find the correlation dimension a better exponent to use for classification of bacteria than the Hurst exponent in this level.

ACKNOWLEDGEMENTS

One of the authors Zu-Guo Yu would like to express his thanks to Prof. Bai-lin Hao of Institute of Theoretical Physics of Chinese Academy of Science for introducing him into this field and continuous encouragement. He also wants to thank Dr. Bin Wang of ITP and Dr. Fawang Liu at QUT for useful discussions about computer programmes. This project is supported by Postdoctoral Research Support Grant No. 9900658 of QUT.

References

[1] Liaofu Luo, Weijiang Lee, Lijun Jia, Fengmin Ji and Lu Tsai, *Phys. Rev. E* **58(1)** (1998) 861-871.

[2] W. Li and K. Kaneko, *Europhys. Lett.* **17** (1992) 655.

[3] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, *Europhys. Lett.* **23** (1993) 373.

[4] (a) R. Voss, *Phys. Rev. Lett.* **68** (1992) 3805; (b) *Fractals* **2** (1994) 1.

[5] C.K. Peng, S. Buldyrev, A.L.Goldberg, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley, *Nature* **356** (1992) 168.

[6] H.E. Stanley, S.V. Buldyrev, A.L. Goldberg, Z.D. Goldberg, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.K. Peng, and M. Simons, *Physica A* **205** (1994) 214.

[7] H.Herzel, W. Ebeling, and A.O. Schmitt, *Phys. Rev. E* **50** (1994) 5061.

[8] P. Allegrini, M. Barbi, P. Grigolini, and B.J. West, *Phys. Rev. E* **52** (1995) 5281.

[9] S. V. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley and G.M. Visvanathan, *Physica A* **249** (1998) 430-438.

[10] L. F. Lou, Tsai Li, and Y. M. Zhou, *J. Theor. Biol.* **130** (1988) 351.

[11] L.F. Luo and L. Tsai, *Chin. Phys. Lett.* **5** (1988) 421-424.

[12] Juan Zhen and Zu-Guo Yu, Correlation dimension and Kolomogrov entropy of DNA sequences, *J. of Xiangtan University* **22(1)** (2000) 115-119.

[13] Zu-Guo Yu, Bai-lin Hao, Hui-min Xie and Guo-Yi Chen, Dimension of fractals related to language defined by tagged strings in complete genome. *Chaos, Solitons and Fractals* (2000) (to appear).

[14] Zu-Guo Yu and Guo-Yi Chen, Rescaled range and transition matrix analysis of DNA sequences. *Comm. Theor. Phys.* (2000) **33(4)** (2000) 673-678.

[15] Zu-Guo Yu and Bin Wang, A time series model of CDS sequences on complete genome, *Chaos, Solitons and Fractals* (2000) (to appear).

[16] Ruqun Shen, Rensheng Chen, Lunjiang Lin, Jian Sun, Yi Xiao, and Jun Xu, *Chinese Science Bulletin (in Chinese)* **38** (1993) 1995-1997.

[17] C. M. Fraser *et al.*, The minimal gene complement of Mycoplasma genitalium, *Science*, **270** (1995) 397.

[18] Bai-lin Hao, Hoong-Chien Lee, and Shu-yu Zhang, Fractals related to long DNA sequences and complete genomes, *Chaos, Solitons and Fractals*, **11(6)** (2000) 825-836.

[19] Bai-Lin Hao, Hui-Ming Xie, Zu-Guo Yu and Guo-Yi Chen , Avoided strings in bacterial complete genomes and a related combinatorial problem. *Ann. of Combinatorics.*, to appear (2000).

[20] A. Provata and Y. Almirantis, Fractal Cantor patterns in the sequence structure of DNA. *Fractals* **8(1)** (2000) 15-27.

[21] P. Grassberger and I. Procaccia, *Physica D* **9** (1983) 189.

[22] P. Grassberger and I. Procaccia, *Phys. Rev. Lett.* **50** (1983) 346.

[23] Hao Bai-lin, Elementary Symbolic Dynamics and Chaos in Dissipative Systems. World Scientific, Singapore, 1989.

[24] H.E. Hurst, Long-term storage capacity of reservoirs, *Trans. Amer. Soc. Civ. Eng.* **116** (1951) 770-808.

[25] B.B. Mandelbrot, *The Fractal Geometry of Nature*, W. H. Freeman, New York, 1982.

[26] J. Feder, *Fractals*, Plenum Press, New York, London, 1988.

[27] N. Iwabe *et al*, Evolutionary relationship of archeabacteria,eubacteria, and eukaryotes infer from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA* **86** (1989) 9355-9359.

[28] E. Pennisi, Genome data shake the tree of life, *Science* **286** (1998) 672.

[29] K. Lechner, G. Heller & A. Böck, *Nucleic Acids Res.* **16** (1988) 7817-7826.